PEOPLE ANALYTICS

for better business performance

People Analytics

What is it?	Why?	How?
data-driven approach for Human Resource Management	Identify Attract Develop Retain TALENT	Building Statistical Models by Analyzing patterns & causal relations

People analytics = Workforce analytics + HR analytics

Business Understanding

Helps an organization to...

- to understand why best and most experienced employees leave prematurely
- to predict which valuable employees will leave next
- to find out what employees feel about their workplace

Goals

Perform Exploratory Data Analysis (EDA)	to find general behavior of Employees
Discover Insights	to identify valued employees & find factors contributing to attrition
Sentiment Analysis	using text mining approach of Natural Language Processing
Develop Statistical Models	Logistic Regression , Decision trees & Random forest to find best-fit
Deploy Best-Fit Model	to predict who will leave next

Data Preparation

Human Resource Analytics Data from Kaggle

10 Attributes / 14999 Rows

Numeric/Factors/Integers

Data Metrics such as Satisfaction level, Last Evaluation, Number of projects, Time Spent, etc.

Employee Reviews Data from Kaggle

Used 3 Attributes

Text Data/Comments/Opinions

Data Metrics such as Pros, Cons & Advice to Management

Clubbed these two datasets into one dataset

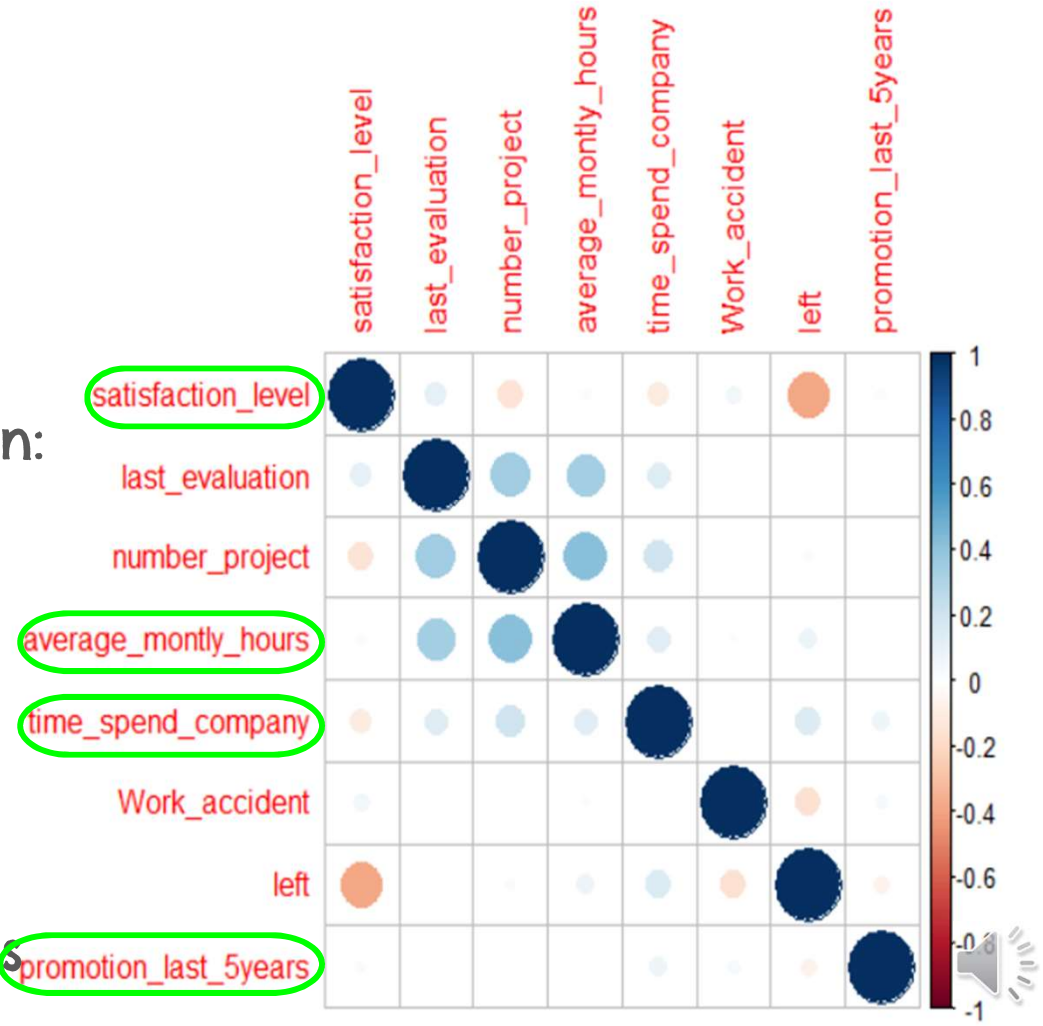
Glimpse of Dataset

```
Console ~/Data Mining/HR Analytics/
> glimpse(hr.data)
Observations: 14,999
Variables: 21
$ satisfaction_level          <dbl> 0.38, 0.80, 0.11, 0.72, 0.37, 0.41, 0.1...
$ last_evaluation             <dbl> 0.53, 0.86, 0.88, 0.87, 0.52, 0.50, 0.7...
$ number_project              <int> 2, 5, 7, 5, 2, 2, 6, 5, 5, 2, 2, 6, 4, ...
$ average_montly_hours        <int> 157, 262, 272, 223, 159, 153, 247, 259,...
$ time_spend_company          <int> 3, 6, 4, 5, 3, 3, 4, 5, 5, 3, 3, 4, 5, ...
$ Work_accident               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ left                        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ promotion_last_5years       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ department                  <fct>  sales ,  sales ,  sales ,  sales ,  sa...
$ salary                      <fct>  low,  medium,  medium,  low,  low,  lo...
$ summary                     <fct> "Best Company to work for", "Moving at ...
$ pros                        <fct> "People are smart and friendly", "1) Fo...
$ cons                        <fct> "Bureaucracy is slowing things down", "...
$ advice.to.mgmt              <fct> "none", "1) Don't dismiss emotional int...
$ overall.ratings             <int> 5, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...
$ work.balance.stars          <fct> 4, 2, 5, 2, 5, 4, 5, 5, 5, 5, 4, 5, 5, ...
$ culture.values.stars        <fct> 5, 3, 4, 5, 5, 4, 4, 5, 5, 5, 5, 5, 5, ...
$ carrer.opportunities.stars  <fct> 5, 3, 5, 5, 5, 4, 4, 5, 5, 5, 4, 5, 5, ...
$ comp.benefit.stars          <fct> 4, 5, 5, 4, 5, 5, 5, 5, 5, 5, 4, 5, 5, ...
$ senior.mangemnet.stars      <fct> 5, 3, 4, 5, 5, 4, 4, 5, 5, 5, 3, 5, 5, ...
$ helpful.count               <int> 0, 2094, 949, 498, 49, 1, 0, 0, 0, 0, 0...
```

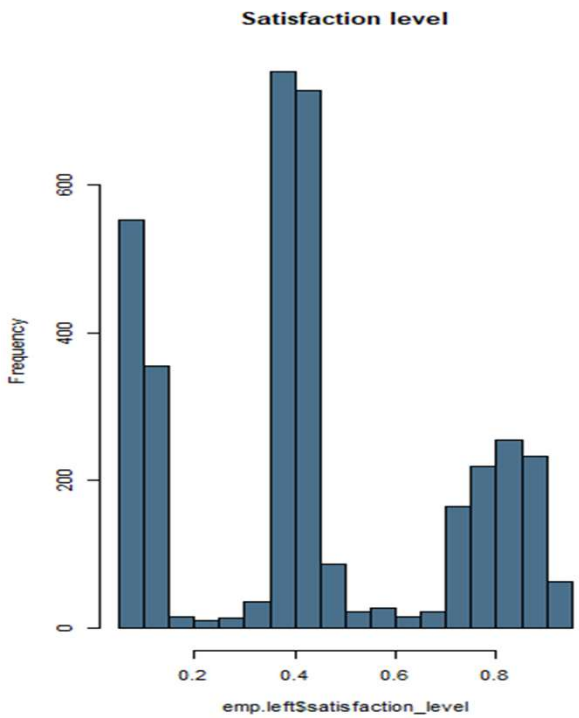
Correlation Matrix

Factors responsible for Attrition:

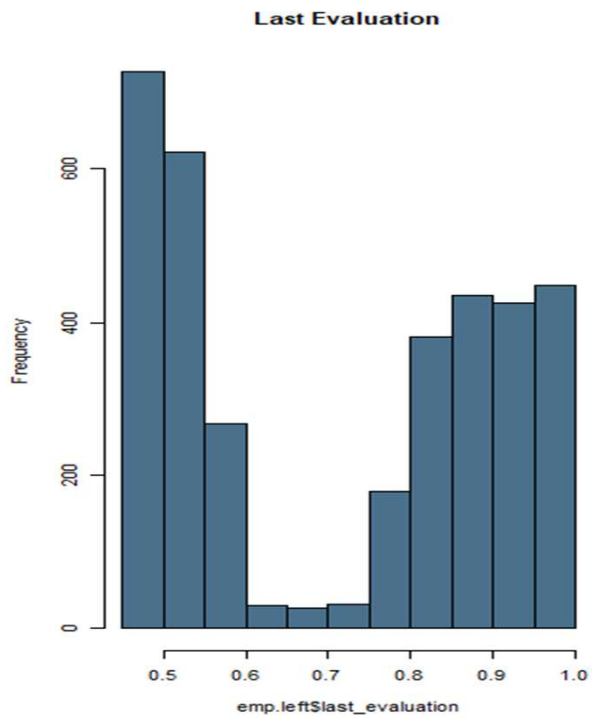
- ❑ low satisfaction level
- ❑ working hours
- ❑ time spend in company
- ❑ promotion within five years



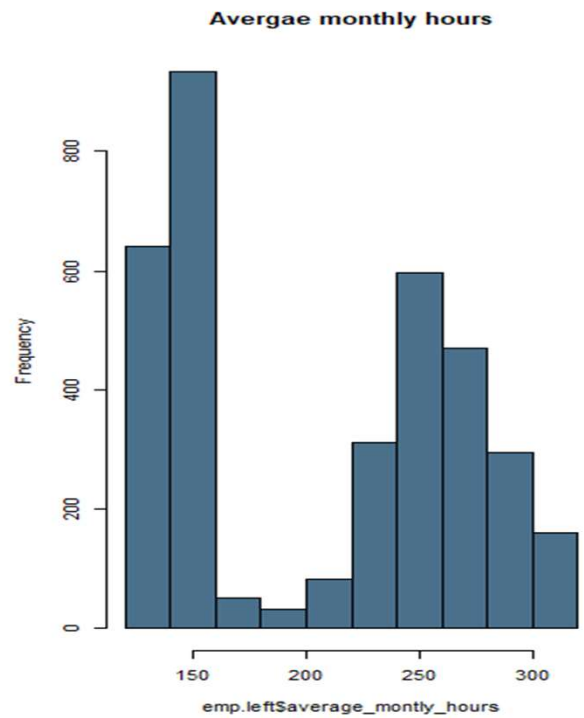
Satisfaction/Evaluation/Working Hours ~ Attrition



Low Satisfaction Level



Low & High Performers



Less & More Working Hours

Attrition Rate vs Department

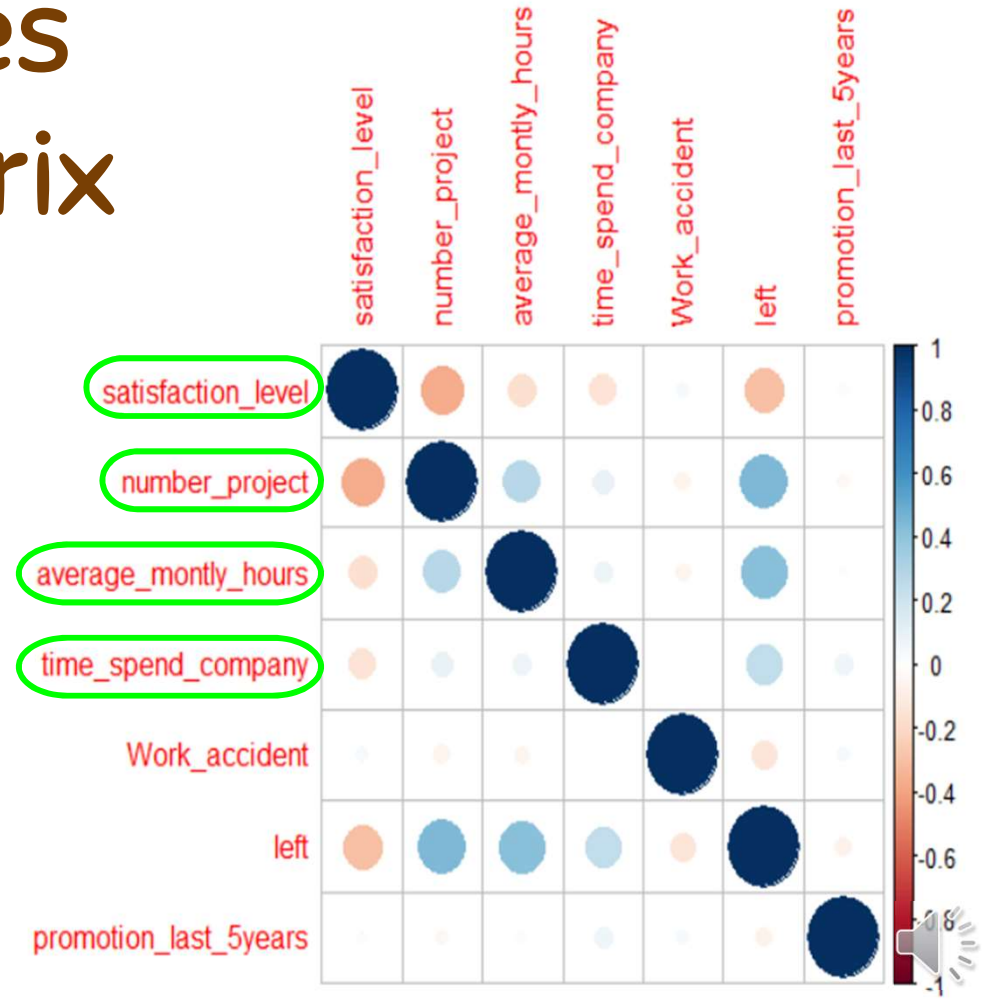
Good Employees Correlation matrix

Criteria for good Employees

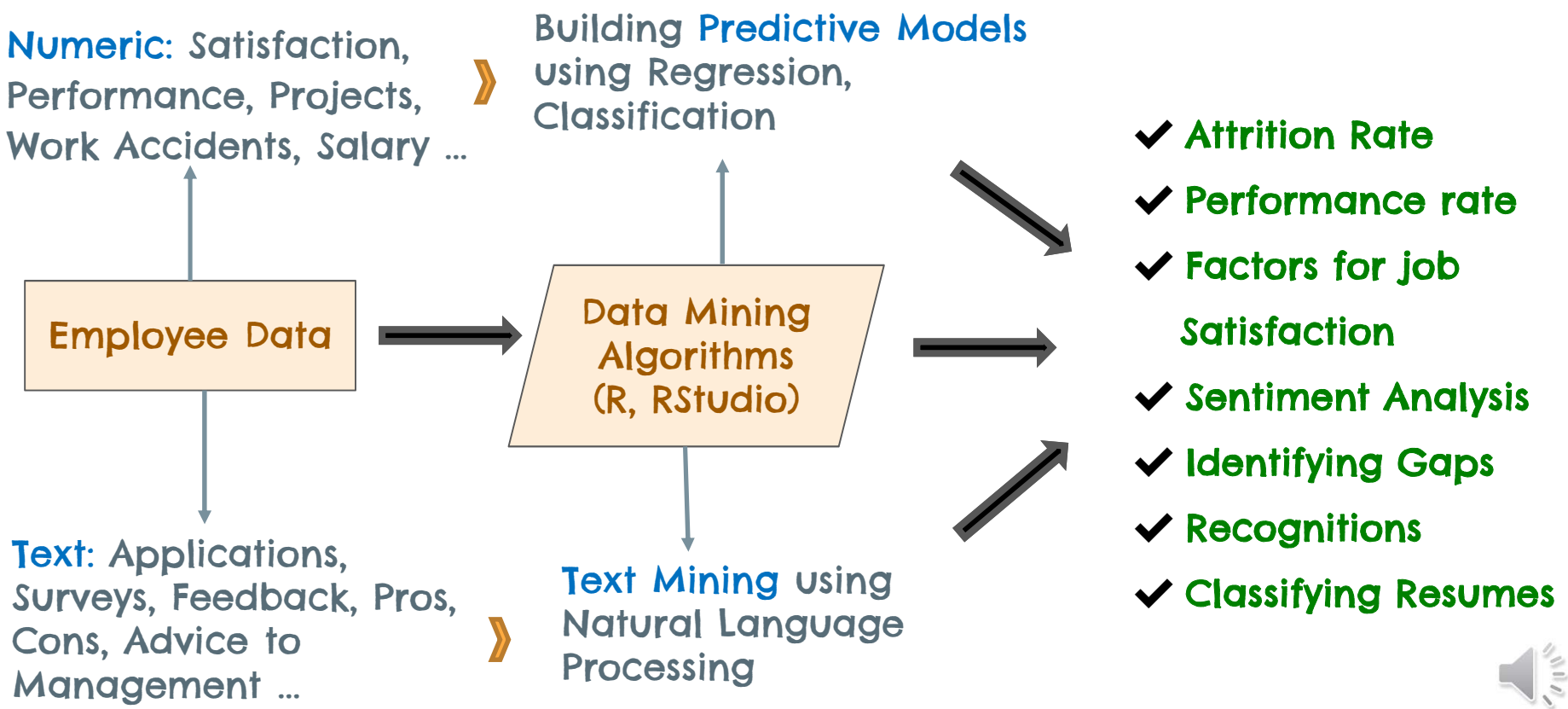
last_evaluation >= 0.70

time_spend_company >= 4

number_project > 5



Experimental Model



Employees Sentiment Analysis

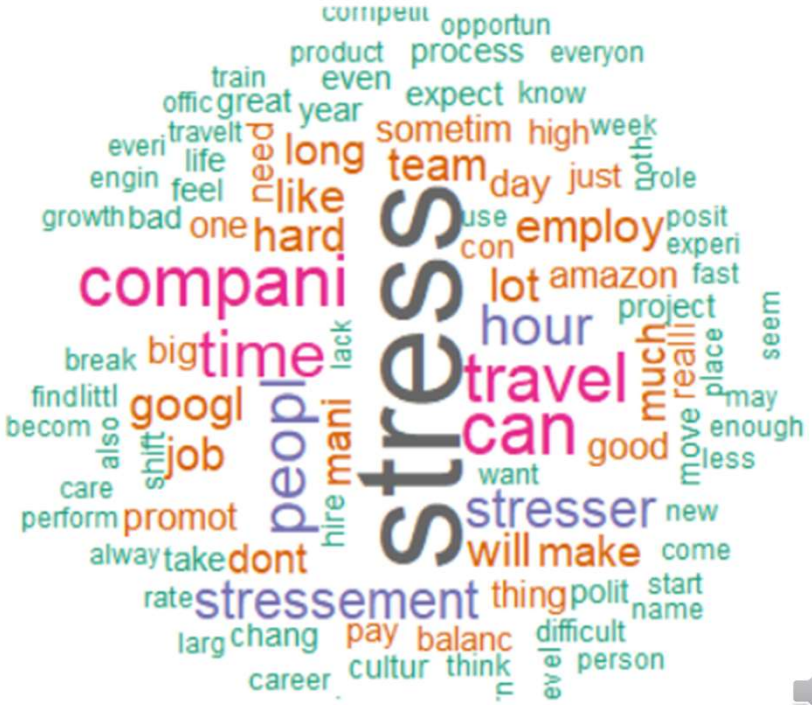
Performed Text Mining to get Insights from Reviews of Employees

Text Extraction Process for Natural Language Processing:

- Convert into Corpus
- Lower Case
- Remove Punctuation
- Remove Stop Words
- Stemming
- Convert DTM to TDM

Word Cloud for Pros and Cons Reviews

```
# Create a wordcloud for the values in word_freqs
wordcloud(pros_word_freq$term,pros_word_freq$num,max.words=100,random.order=F,colors=brewer.pal(8,"Paired"))
wordcloud(cons_word_freq$term,cons_word_freq$num,max.words=100,random.order=F,colors= brewer.pal(8,"Dark2"))
```



Logistic Regression

```
# logistic regression model
modelglm <- train(left ~ . - department
                  - employeeNumber
                  - Work_accident,
                  data=emp.good,
                  trControl=train_control,
                  method='glm',
                  family='binomial')
```

Training the Model

Predictions

	Remained	Left
Predicted to stay	9734	628
Predicted to leave	301	1419

```
Console ~/Data Mining/HR Analytics/
> cMatrixglm<- confusionMatrix(predsglm, emp.good$left)
> cMatrixglm
Confusion Matrix and Statistics

          Reference
Prediction Remained Left
  Remained     9735  599
  Left          300 1448

               Accuracy : 0.9256
                 95% CI : (0.9208, 0.9302)
    No Information Rate : 0.8306
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7193
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9701
            Specificity : 0.7074
         Pos Pred Value : 0.9420
         Neg Pred Value : 0.8284
             Prevalence : 0.8306
         Detection Rate : 0.8057
   Detection Prevalence : 0.8553
      Balanced Accuracy : 0.8387

       'Positive' Class : Remained
```

Confusion Matrix

92.5% Accuracy

Decision Trees

```
# Decision Tree model
modelrpart<- train(left~. - employeeNumber,
                   data=emp.good,
                   trControl=train_control,
                   method="rpart")
```

Training the Model

Predictions

	Remained	Left
Predicted to stay	9868	308
Predicted to leave	167	1739

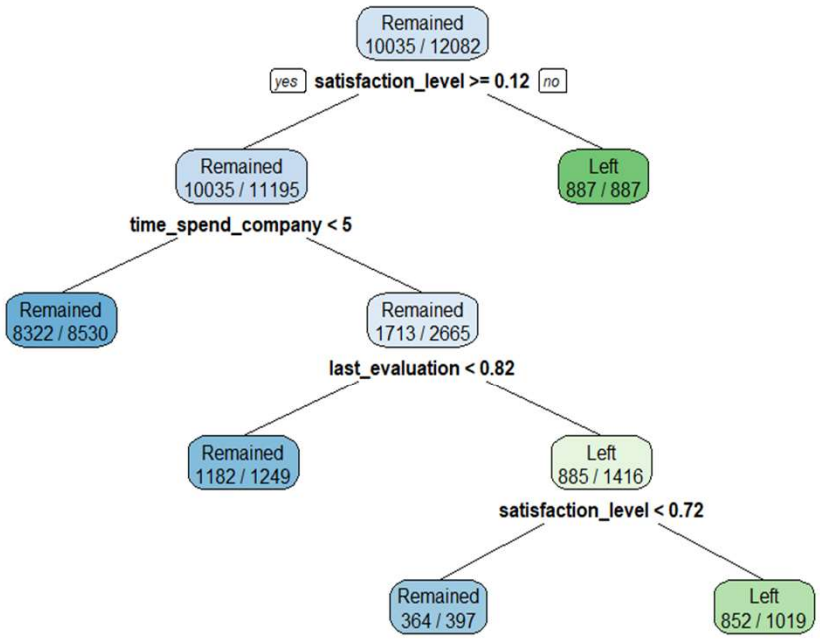
```
Console ~/Data Mining/HR Analytics/
> # Confusion Matrix for Decision Tree Model
> cMatrixrpart
Confusion Matrix and Statistics

          Reference
Prediction Remained Left
  Remained     9868  308
  Left          167 1739

               Accuracy : 0.9607
                 95% CI : (0.9571, 0.9641)
    No Information Rate : 0.8306
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8564
 Mcnemar's Test P-Value : 1.331e-10

            Sensitivity : 0.9834
            Specificity : 0.8495
         Pos Pred Value : 0.9697
         Neg Pred Value : 0.9124
             Prevalence : 0.8306
         Detection Rate : 0.8168
   Detection Prevalence : 0.8422
      Balanced Accuracy : 0.9164

       'Positive' Class : Remained
```

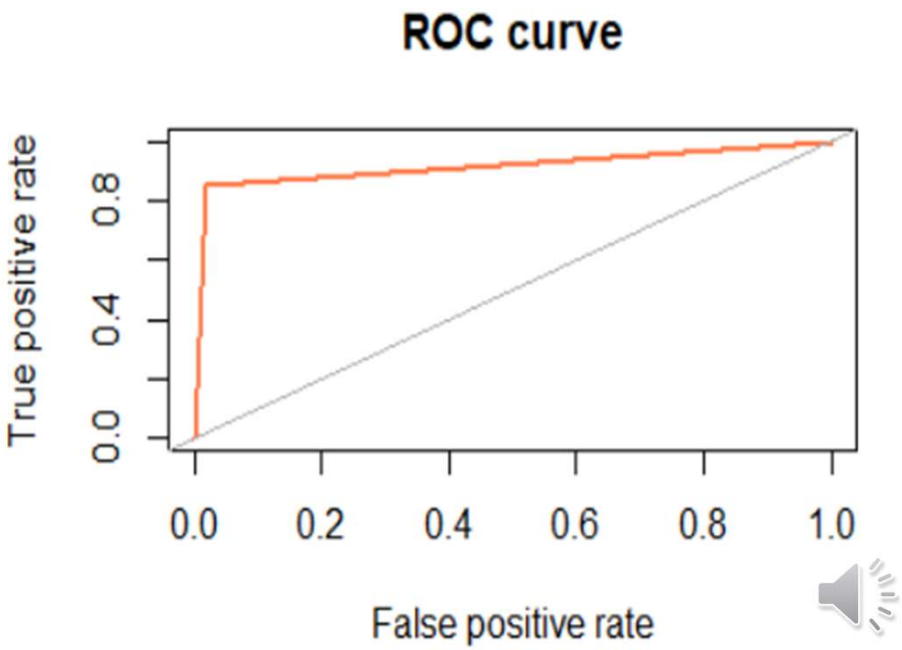
Confusion Matrix

96% Accuracy

Plotting Trees & ROC Curve

```
# plotting the decision trees
rpart.plot(modelrpart$finalModel, type = 2, fallen.leaves = F, cex = 1, extra = 2)
```





Random Forest

```r
# Random Forest Model
modelrf <- train(as.factor(left)~.,
                 data=emp.good,
                 method="rf",
                 metric="Accuracy",
                 tuneGrid=tunegrid,
                 trControl=train_control)
```

Training the Model

Predictions

	Remained	Left
Predicted to stay	10026	163
Predicted to leave	9	1884

```
Console ~/Data Mining/HR Analytics/
> confusionMatrix(emp.good_rf_pred, as.factor(emp.good$left))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4956   60
         1    9 1840

               Accuracy : 0.9899
                 95% CI : (0.9873, 0.9922)
    No Information Rate : 0.7232
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9747
 Mcnemar's Test P-Value : 1.752e-09

            Sensitivity : 0.9982
            Specificity : 0.9684
         Pos Pred Value : 0.9880
         Neg Pred Value : 0.9951
             Prevalence : 0.7232
         Detection Rate : 0.7219
   Detection Prevalence : 0.7307
      Balanced Accuracy : 0.9833

       'Positive' Class : 0
```

Confusion Matrix

99% Accuracy

Random Forest Deployment

```
leavingProbabilityglm <- predict(modelrf, emp.good, type='prob')
datatable(glmProbTable, options = list(
        columnDefs = list(list(class Name = 'dt-center', targets = c(1, 4, 5, 6)))
```

Show 10 entries Search:

	employeeNumber	department	salary	probLeaving	last_evaluation
1	6727	sales	low	0.954	0.95
2	2704	hr	low	0.926	0.93
3	2571	support	medium	0.916	0.96
4	3080	sales	medium	0.902	0.97
5	7224	support	medium	0.874	0.98
6	7356	sales	medium	0.814	0.94
7	3781	technical	low	0.792	0.96
8	6359	sales	medium	0.786	0.98
9	10099	sales	low	0.782	0.83
10	11135	technical	medium	0.782	0.85

Showing 1 to 10 of 200 entries Previous 1 2 3 4 5 ... 20 Next

Take Home...

Pros:

Data Mining on HR datasets will help data driven decision making

We can identify employee sentiments, and explore how results of EDA and Data Mining Algorithm turn into actionable Key Performance Indicators (KPI) for Human Resource Management

Cons:

According to recent survey, HR do not want the model to make decisions instead they want a model which can assist them in decision making process for retaining good employees having higher probability-to-leave

The Team

Aakash Sarap

Master's Degree in Analytics
Northeastern University

Sushmita Jadhav

Master's Degree in Analytics
Northeastern University